NOTRE DAME MATHEMATICAL LECTURES
Number 1

ON THE PRINCIPLES OF STATISTICAL INFERENCE

Four Lectures

Delivered at the University of Notre Dame
February 1941

by

DR. ABRAHAM WALD
Professor, Columbia University

NOTRE DAME, INDIANA
1942

Lithoprinted in U.S.A.
EDWARDS BROTHERS, INC.
ANN ARBOR, MICHIGAN

I INTRODUCTION

The purpose of statistics, like that of geometry or physics, is to describe certain real phenomena. The objects of the real world can never be described in such a complete and exact way that they could form the basis of an exact theory. We have to replace them by some idealized objects, defined explicitly or implicitly by a system of axioms. For instance, in geometry we define the basic notions "point," "straight line," and "plane" implicitly by a system of axioms. They take the place of empirical points, straight lines and planes which are not capable of exact definition. In order to apply the theory to real phenomena, we need some rules for establishing the correspondence between the idealized objects of the theory and those of the real world. These rules will always be somewhat vague and can never form a part of the theory itself.

The purpose of statistics is to describe certain aspects of mass phenomena and repetitive events. The fundamental notion used is that of "probability." In the theory it is defined either explicitly or implicitly by a system of axioms. For instance, Mises[1] defines the probability of an event as the limit of the relative frequency of this event in an infinite sequence of trials satisfying certain conditions. This is an explicit definition of probability. Kolmogoroff[2] defines probability as a set function which satisfies a certain system

---

1) See references 10 and 11

2) See reference 9

of axioms. These idealized mathematical definitions are related to the applications of the theory by translating the statement "the event E has the probability p" into the statement "the relative frequency of the event E in a long sequence of trials is approximately equal to p." This translation of a theoretical statement into an empirical statement is necessarily somewhat vague, for we have said nothing about the meanings of the words "long" or "approximately." But such vagueness is always associated with the application of theory to real phenomena.

It should be remarked that instead of the above translation of the word "probability" it is satisfactory to use the following somewhat simpler one: "The event E has a probability near to one" is translated into "it is practically certain that the event E will occur in a single trial." In fact, if an event E has the probability p then, according to a theorem of Bernoulli, the probability that the relative frequency of E in a sequence of trials will be in a small neighborhood of p is arbitrarily near to 1 for a sufficiently long sequence of trials. If we translate the expression "probability nearly 1" into "practical certainty," we obtain the statement "it is practically certain that the relative frequency of E in a long sequence of trials will be in a small neighborhood of p."

In statistics we always construct some probability schemes which we believe to be adequate to describe certain real phenomena. For instance, we describe the situation concerning the possible outcomes in tossing a coin by saying that the probability of obtaining a head in one toss is 1/2, for in a long se-

quence of trials we would expect to have about half as many heads as total tosses. Or, if we measure the length of a bar by some instrument, we sometimes assume that the result is a normally distributed random variable. The notions of a random variable and a distribution function are defined as follows: if $F(x)$ is a function expressing the probability that a real variable $X < x$, we say that X is a random variable and that $F(x)$ is the probability distribution of X. Then, if $F(x)$ is given by the formula

$$(1) \qquad F(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{\frac{-1}{2}\frac{(y-\mu)^2}{\sigma^2}}\, dy$$

we say that X is normally distributed. The quantities σ and μ are real parameters. Thus, if in measuring the length of a bar by some instrument we assume that the outcome of the measurement is a normally distributed random variable, we may express the probability that a measurement will be less than a given value x by (1).

If $X_1, X_2, X_3, \ldots, X_n$ represent n random variables and $x_1, x_2, \ldots, x_n$ any set of real numbers, we use the symbol $F(x_1, x_2, \ldots, x_n)$ to express the probability of the composite event that $X_1 < x_1$, $X_2 < x_2, \ldots, X_n < x_n$ simultaneously. This function will be called the joint probability distribution of the n random variables. We shall say that n random variables are independently distributed if the function $F(x_1, x_2, \ldots, x_n)$ is the product of n functions such that only x_1 is involved in the first, only x_2 in the second, and so on. That is

$$F(x) = f_1(x_1) f_2(x_2) \ldots f_n(x_n).$$

For example, if n measurements X_1, X_2,..., X_n of a bar are independently and normally distributed with the same normal distribution, we would obtain

(2) $$F(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^2} \int_{-\infty}^{x_1} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy \int_{-\infty}^{x_2} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy \cdots \int_{-\infty}^{x_n} e^{\frac{-(y-\mu)^2}{2\sigma^2}} dy.$$

If we measure the length of a bar n times by some instrument, we sometimes find it appropriate to adopt the probability scheme that the results of the n measurements have a joint probability distribution given by (2).

One of the fundamental problems of statistical inference is that of testing statistical hypotheses. The most general form of a statistical hypothesis we have to deal with in statistical theory may be expressed as follows. Let $X_1,\ldots,X_n$ be a finite set of random variables and let $F(x_1,\ldots,x_n)$ be its joint probability distribution function. Then the statistical hypothesis is the statement that the unknown distribution function $F(x_1,\ldots,x_n)$ is an element of a certain class ω of distribution functions. For instance, if $X_1,\ldots X_n$ are successive measurements on the length of a bar, we may consider the hypothesis that $X_1,\ldots,X_n$ are independently distributed with the same normal distribution. In this case ω is a two parameter family given by (2), σ being any positive number and μ any real number.

If we consider the hypothesis that $X_1,\ldots,X_n$ are normally, independently distributed with zero means ($\mu=0$) and unit variances ($\sigma^2=1$), then ω consists of a single element. When the class ω consists of a single element, we shall say that the hypothesis we are considering is a simple hypothesis. Otherwise, it will be called composite.

The question of testing a given hypothesis may be formulated in the following manner. We should like to know, on the basis of n observations $x_1,\ldots,x_n$ where x_α is the observed value of the random variable X_α ($\alpha=1,\ldots,n$), whether to accept or reject the hypothesis H_ω that the unknown distribution function $F(x_1,\ldots,x_n)$ belongs to the class ω. The set of n observations can be represented by a point E of n-dimensional Cartesian space, called the sample space. To test the hypothesis H_ω on the basis of n observations we must choose a subset R of the sample space and then reject the hypothesis H_ω if the sample point E falls within R. Otherwise, we maintain the hypothesis. It is evident that the fundamental problem here is the choice of the subset R, which we shall call the critical region. The solution of this problem depends, to some extent, upon any a priori knowledge we may have about the unknown distribution function $F(x_1,\ldots,x_n)$. One of the most important and most frequent a priori assumptions is that the random variables $X_1,\ldots,X_n$ are independently distributed, each having the same distribution. Thus, we have the assumption that F is of the form

$$F(x_1,\ldots,x_n) = \prod_1^n \varphi_i(x_i) \quad \text{where } \varphi_i = \varphi_j \text{ for all } i,\ j.$$

Such a priori knowledge about our unknown distribution function can always be expressed by saying that the function

$F(x_1,\ldots,x_n)$ is an element of a certain class Ω of distribution functions. The class ω which is being considered is then always a subclass of Ω. We shall see that the choice of the critical region R for testing the hypothesis H_ω will depend upon the a priori knowledge Ω.

It is now seen that the problem of testing hypotheses can be formulated as follows: Taking for granted that the unknown distribution function F is an element of a class Ω, we wish to test the hypothesis that F belongs to a certain subclass ω of Ω. The problem to be solved is the question of how the critical region in the sample space should be chosen.

For instance, Ω may be defined by the statement that $X_1,\ldots,X_n$ are independently and normally distributed each of them having the same distribution, and ω may be the subclass of Ω defined by the additional restriction that the mean values of $X_1,\ldots,X_n$ are zero. In this case, according to certain standards we will discuss later, the adequate critical region is given by the inequality

$$\left|\frac{\bar{x}\sqrt{n}}{s}\right| \geq c$$

where $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ and $s^2 = \frac{\sum_{\alpha=1}^{n}(x_\alpha-\bar{x})^2}{n-1}$

and c is a certain constant. If, however, Ω is a much broader class defined by the statement that $X_1,\ldots,X_n$ are independently distributed each having the same distribution, the above critical region for testing H_ω is not adequate, and some other critical region has to be chosen.

Before we proceed farther it might be well for us to list a few of the mathematical terms used together with their meanings in statistics. We can do this in tabular form.

MATHEMATICAL TERMINOLOGY	STATISTICAL INTERPRETATION
n space, E_n (sample space)	Possible outcome of n observations.
Ω, class of functions on E_n	Class of possible probability distributions.
ω, subclass of Ω	The statistical hypothesis. The true distribution is a member of ω.
R, (critical region), a subset of E_n	Criterion for rejecting the hypothesis that the true distribution is a member of ω.
Association of R with Ω and ω.	Choice of the critical region for testing the hypothesis.

The problem of testing hypotheses is only one of the problems of statistical inference. Another is the problem of estimation. Given that the unknown distribution function F belongs to a certain class Ω of distribution functions, how can we choose a function $\varphi(E)$ defined for all points E of E_n such that the value of $\varphi(E)$ is always an element of Ω and can be considered a "good" estimate of the unknown distribution function F? We may say that $\varphi(E)$ is a "good statistical estimate" of F if the probability is as large as possible that $\varphi(E)$ is in a small neighborhood of F. We will formulate this principle more precisely in chapter III.

If, for instance, Ω is given by the statement that $X_1,\ldots,X_n$ are independently and normally distributed with the same means and unit variances, then Ω is a one parameter family of distribution functions and an element of Ω is completely specified by specifying the value of the unknown mean μ.

Hence, to estimate the unknown distribution function F is the same as to estimate the unknown mean μ. In this case the problem of estimation is the problem of finding a real function $\varphi(E)$ defined for all points E of the sample space such that $\varphi(E)$ can be considered as a statistical estimate of the unknown mean μ. The classical solution of this problem in this particular case is given by

$$\varphi(E) = \frac{x_1+\ldots+x_n}{n} .$$

The two types of problems of statistical inference mentioned so far do not cover all possible problems.[3] The following problem, for example, is neither a problem of testing a hypothesis nor one of estimation: Consider three subclasses ω_1, ω_2, ω_3 of the class Ω of distribution functions, and denote by H_{ω_1} the hypothesis that the unknown distribution F is an element of ω_1. The problem considered is to decide on the basis of the n observations which of the three hypotheses should be accepted (assume that the sum of the three subclasses ω_1, ω_2, ω_3 is equal to Ω). Such a situation may arise, for instance, in the case of a manufacturer who has to keep the quality of his product between two limits, and wants to test, by sampling, whether the quality is actually between these limits, below the lower limit, or above the upper limit. (Assume that the quality is measurable and can be represented by a real number.)

---

3) See in this connection 16, pp 299-300.

The reasons why such a "trilemma" is a problem different from testing a hypothesis or estimation can only be indicated here. It will be seen that there are many approaches to each problem of inference, and that the theory provides means of choosing among them by deciding that certain approaches are "better" than certain others. Now, one might suggest the reduction of the above "trilemma" to a problem of, say, estimation by estimating the unknown distribution function F and accepting that hypothesis which corresponds to the subclass in which the estimate of F is contained. This would be one answer to the trilemma, but by no means the "best" answer according to the standards developed.

The most general formulation of the problem of statistical inference is this: Let S be a system of subclasses of the class Ω of distribution functions. For each element s of S, consider the hypothesis H_s which states that the unknown distribution F is an element of s; denote by H_S the system of all such hypotheses; the problem is to decide, by means of a sample, which element of H_S should be accepted.

The problems enumerated before are special cases of this general problem. If S consists of two elements only, one being a subclass ω of Ω and the other its complement in Ω, the problem is the same as that of testing the hypothesis that the true distribution function F is an element of ω. If S is the system of all elements of Ω, we have the problem of estimation. If S consists of three classes ω_1, ω_2, ω_3 with the sum Ω, we have the trilemma.

II THE NEYMAN-PEARSON THEORY OF TESTING A STATISTICAL HYPOTHESIS [4]

The principles of statistical inference as developed in the last two decades by R.A.Fisher, Neyman and Pearson deal with the problem of testing a hypothesis and with the problem of estimation but not with the general problem of statistical inference as it has been formulated in the foregoing pages. A further restriction in these theories is that they deal only with the case that Ω is a k-parameter family of distribution functions, i.e., that the true but unknown distribution function F is known to be an element of a k-parameter family of functions

$$F(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_k)$$

where $\theta_1, \ldots, \theta_k$ are parameters. In this case the specification of the values of the parameters specifies completely the distribution function F.

A set of parameter values can be represented by a point in a k-dimensional Euclidean space called a parameter space. Because of the one-to-one correspondence between elements of Ω and points of the parameter space we can identify Ω with the parameter space. If for example, $X_1, \ldots, X_n$ are normally and independently distributed, each having the same distribution (equation(2)), then the parameter space is a half plane where $\theta_1 = \mu$ = mean value, and $0 < \theta_2 = \sigma$ = standard deviation.

A hypothesis concerning F is expressed by the statement that the true parameter point lies in a certain subset ω of the parameter space Ω. As we have done before, we shall call the hypothesis a simple one if ω consists of a single point.

4) See, in this connection, references 12,13 and 14

Otherwise, it is called a composite hypothesis. In the above example the statement that $\mu = 0$, $\sigma = 1$ is a simple hypothesis, while merely stating that $\mu = 0$ without specifying σ is a composite hypothesis.

For the sake of simplicity we shall confine ourselves to the case of a single unknown parameter since this suffices to illustrate the basic ideas of the theories of Fisher, Neyman and Pearson. First, we shall deal with the Neyman-Pearson theory of testing a statistical hypothesis.

We assume that the unknown distribution function is known to be an element of a one-parameter family $F(x_1, x_2, \ldots, x_n, \theta)$ and we wish to test the hypothesis $\theta = \theta_0$.

A simple example for this case is the following: Let it be known that $X_1, \ldots, X_n$ are independently and normally distributed with the same mean and unit variances, i.e., Ω is the one-parameter family of distributions

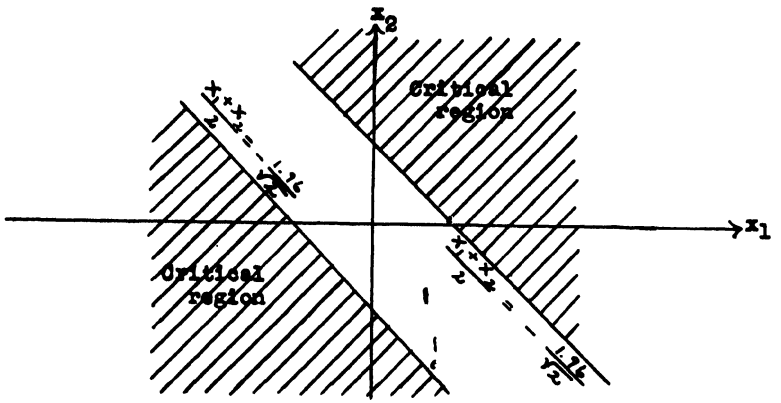
$$F(x_1, \ldots, x_n, \theta) = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{x_1} e^{\frac{-(v-\theta)^2}{2}} dv \ldots \int_{-\infty}^{x_n} e^{\frac{-(v-\theta)^2}{2}} dv,$$

and assume that we wish to test the hypothesis that $\theta = 0$. According to the classical theory we reject this hypothesis if and only if

$$\left|\bar{x}\right| \geq c; \quad \left(\bar{x} = \frac{x_1 + \ldots + x_n}{n}\right)$$

where c denotes a certain constant. The value of c is chosen in such a way that the probability of $\left|\bar{x}\right| > c$ under the assumption that the hypothesis $\theta = 0$ is true, is so small that we are willing to reject the hypothesis. If we want this probability to be 5 percent, then $c = \frac{1.96}{\sqrt{n}}$.

If, in the same example, we have made only two observations x_1, x_2, so that the sample space is the Euclidian plane, the critical region consists of all points for which $\frac{1}{2}(x_1+x_2) > \frac{1.96}{\sqrt{2}}$ and all points for which $\frac{1}{2}(x_1+x_2) < \frac{-1.96}{\sqrt{2}}$. If the point representing the observations falls within the critical region (i.e., if the arithmetic mean of the two observations is larger than $\frac{1.96}{\sqrt{2}}$ or smaller than $\frac{-1.96}{\sqrt{2}}$) we shall reject the hypothesis that the mean value is zero.



But the classical theory does not suggest why this critical region should be used. It merely proves that the probability for the observation point to fall within the critical region is five percent when the initial hypothesis is fulfilled. But there are, infinitely many regions which enjoy the same property, and the classical theory does not give any reasons why just the one region mentioned should be chosen.

In order to arrive at a distinction, between various critical regions, Neyman and Pearson advance the following considerations. In making a statement of acceptance or rejection of a

hypothesis, we may commit two types of errors: rejecting the hypothesis although it is true (error of type I), or failing to reject it although it is false (error of type II). If the hypothesis consists in saying that the unknown parameter θ has a given value θ_0, the situation may be summarized as follows:

Truth or Falsehood of Statement Concerning the Hypothesis $\theta = \theta_0$

True Situation	Statement Advanced	
	$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta = \theta_0$	Correct	Type I error
$\theta \neq \theta_0$	Type II error	Correct

By size of the critical region we mean the probability that the point representing the observations will fall within the critical region, where the probability in question is calculated under the assumption that the hypothesis is true. (Thus, in the example used before, the size of the critical region was five percent.) This may be expressed by saying that the size of the critical region is equal to the probability of committing a type I error.

The general idea underlying the theory of Neyman and Pearson is to minimize the probability of type II errors while keeping the probability of type I errors constant.

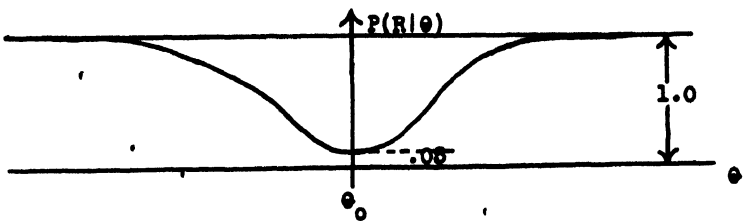
If R is any region in the sample space, and E is the point of the sample space which represents the observations, we shall denote by $P(R|\theta_1)$ the probability of E lying in R calculated

under the assumption that θ_1 is the true value of the unknown parameter θ, that is to say, $P(R|\theta_1)$ is equal to the Stieltjes integral $\int_R dF(x_1,\ldots,x_n, \theta_1)$ over the region R. Thus, if we make the hypothesis $\theta = \theta_0$ and choose R as a critical region for this hypothesis, the size of the critical region will be given by the expression $P(R|\theta_0)$. If the hypothesis is wrong and the true value of θ is θ_1, then the probability of avoiding an error of type II is $P(R|\theta_1)$.

The expression $P(R|\theta_1)$, i.e., one minus the probability of an error of type II, is called the power of the critical region R with respect to the alternative hypothesis $\theta = \theta_1$.

The expression $P(R|\theta)$ is a function of θ. It may be plotted as a curve, the ordinate of which is equal to the size of R if the abscissa is θ_0, and equal to the power of R with respect to the alternative $\theta = \theta_1$ if the abscissa is any value $\theta_1 \neq \theta_0$. This curve is called the power curve of the region R.

In the former example, in which the distribution was normal with unknown mean and unit variance, and the critical region chosen was $|\bar{x}| > \frac{1.96}{\sqrt{n}}$ (where $\bar{x}$ is the arithmetic mean of the observations $x_1, x_2, \ldots, x_n$), the power curve can easily be calculated and has the form shown below:



In order to compare the test $|\bar{x}| > \frac{1.96}{\sqrt{n}}$ with other possible tests, we have to compare the above power curve with the power curves of other critical regions which have the same size, five percent.

In general, if we have two critical regions R and R', both of which have the desired size, and if the power curve of R' is above that of R for the value $\theta = \theta_1$, then the critical region R' is better than R for testing the hypothesis if the true value of θ happens to be θ_1. For the probability of committing a type I error is the same whether R or R' is used, while the probability of committing a type II error when using R' is smaller than when using R. If the power curve of R' is above that of R for each θ (except θ_0 for which the two curves coincide by assumption), then R' will be called uniformly more powerful than R. The test using the critical region R is called non-admissible because its use is, under all circumstances, less favorable than the use of R'.

In order to make this clear, let us assume that a large number of samples is drawn, each of which consists of N individual observations. Let M be the number of such samples and let two statisticians, whom we will call S and S', test the same hypothesis, using each of the M samples. Assume that S uses the critical region R for testing while S' bases his tests on the region R'. S and S' will each obtain M answers to the question as to whether the null hypothesis (the hypothesis to be tested) should be rejected. Some of these answers will be right, others will be wrong. Let us compare the records of S and S'. We have to distinguish between the case that the null hypothesis is true and the case that it is false. a)In the first case, the answers

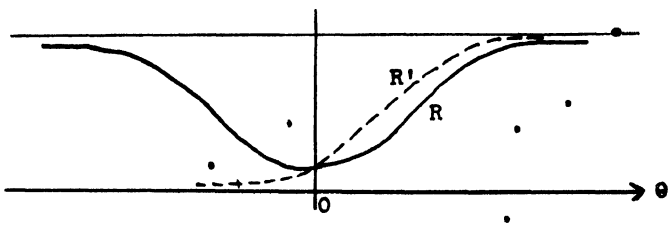
obtained by each statistician may either be that the hypothesis is to be accepted - these answers are right; or that it should be rejected - these answers are errors of type I. The probability of committing a type I error by testing the null hypothesis from a sample drawn at random is equal to the size of the critical region used in testing. If M is large, it is practically certain that the relative frequency of type I errors will be approximately equal to their probability, i.e., to the size of the critical region. Since R and R' have, by assumption, equal size, each of the two statisticians will commit approximately the same number of errors. b)If the null hypothesis is false, some of the M answers obtained by each statistician will correctly reject it, while others will accept it, thus committing errors of type II. If M is large, the relative frequency of correct answers will be approximately equal to the power of the test used which we have pointed out is the probability of avoiding a type II error. By assumption, the power of R' is greater than that of R, regardless of what the true value of θ is, provided only that θ is different from θ_0. Therefore, the relative frequency of wrong answers obtained by S will tend to be greater than the relative frequency of wrong answers obtained by S'. Thus, if the null hypothesis is false (no matter what the true value of θ is), it is practically certain that S will make more false statements; while if the null hypothesis is true, S and S' will commit an approximately equal number of false statements. The method used by S', i.e., the application of the critical region R', is therefore superior to the method used by S, i.e., the application of the critical region R.

These considerations decide the choice between two critical regions of equal size if one of them is uniformly more powerful than the other, i.e., if the power curve of the former is above that of the latter for all values of θ except θ_0 (for which the power curves coincide). On the other hand, if the power curve of R' is above that of R for some values of θ, but below it for other values of θ, then we cannot choose one of the two regions without introducing further principles on which to base the choice.

If, for all values of θ, the power curve of a region R is never below that of any other region R' of equal size, then R is called a uniformly most powerful region, and the test corresponding to R a uniformly most powerful test.

The first principle for selecting a test is this: whenever we can find a uniformly most powerful test, we shall prefer it to all other tests using regions of the same size. Unfortunately, uniformly most powerful tests do not exist in most cases.

In the example which we have used on page 11 let us consider the region R' determined by the inequality $\bar{x} > \frac{1.64}{\sqrt{n}}$. It can easily be shown that R' (like the region R considered before) has the size .05. The power curves of R and R' are shown below:



We can see that for all $\theta > 0$, R' is more powerful than R, and vice versa for $\theta < 0$. In such cases further principles have to be formulated on which the choice should be based. It is clear that the choice we make will depend on our a priori degree of belief in the truth of the different possible values of θ. For instance, if we know a priori that θ cannot be negative, then we shall prefer R'

Moreover, it can be shown that R' is uniformly most powerful if the parameter space is restricted to non-negative values of θ. If negative and positive values of θ are considered a priori as equally possible we will most likely prefer R to R'.

This example shows also that the choice of the critical region depends essentially on Ω. If Ω consists of all non-negative values of θ then the region R' is a uniformly most powerful test. If Ω consists of all non-positive values θ, then the region R'' given by $\bar{x} < \frac{-1.64}{\sqrt{n}}$ is a uniformly best region. Finally, if Ω consists of all real values θ, then the use of the region R seems to be more reasonable than that of R' or R''.

Since uniformly most powerful regions rarely exist, Neyman and Pearson introduced a further principle on which the choice of the critical region should be based, namely, the principle of unbiasedness. A test is called <u>unbiased</u> if the power function of the test has a relative minimum at the value $\theta = \theta_0$ where θ_0 is the hypothesis to be tested.

Some rationalization of this principle can be given: Suppose a test is biased, then for some value θ_1, in the neighborhood of θ_0, the power of the test is less than the size of the region. But this means that the probability of rejecting the hypothesis $\theta = \theta_0$ is larger if θ_0 is true than if θ_1 is true,

which is not a desirable situation.

In general, an infinity of unbiased tests exist, hence we need a further principle in order to select a proper test from among them. We define as a uniformly most powerful unbiased test one which is at least as powerful or more powerful, with respect to all alternate hypotheses, than any other unbiased region of equal size. If a uniformly most powerful unbiased test exists, and if we accept the principle of unbiasedness, then it is obvious that it is the most advantageous test to use. Neyman and Pearson called a critical region corresponding to a uniformly most powerful unbiased test a critical region of type A_1.

Referring to the example previously considered, the critical region given by $|\bar{x}| > c$ is a region of type A_1 for testing the hypothesis in question. Another example of a region of type A_1 is the following: Let $X_1,\ldots,X_n$ be independently and normally distributed with zero means and a common variance. Then, for testing the hypothesis that the common variance σ^2 is equal to σ_o^2, the critical region consisting of all points of the sample space which satisfy at least one of the inequalities

$$x_1^2 + \ldots + x_n^2 > c_1 \quad \text{or} \quad x_1^2 + \ldots + x_n^2 < c_2 ,$$

is a critical region of type A_1 if the constants c_1 and c_2 are properly chosen.

The region of type A_1 exists in an important, but very restricted, class of cases; there are many instances in which it does not exist. Therefore, Neyman and Pearson have introduced a third type of region, known as a region of type A. The region R is said to be of type A if its power function P(W/θ) is

such that

$$1) \quad \left.\frac{\partial P(R|\theta)}{\partial \theta}\right|_{\theta = \theta_0} = 0$$

and

$$2) \quad \left.\frac{\partial^2 P(R|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} \geq \left.\frac{\partial^2 P(R'|\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}$$

for all regions R' which satisfy 1) and have the same size as R. The first condition restricts the region to be unbiased. The second requires the power function of a region of type A to have a greater curvature than that of any other unbiased region of the same size. To put it crudely, it means that the region is most powerful in the neighborhood of θ_0.

A critical region of type A exists under very weak conditions which are fulfilled in most of the practical cases. However, the objection can be raised against a region of type A that we are much more concerned with the behavior of the power function for alternatives θ which are far from θ_0 than for those in the neighborhood of θ_0. In spite of this, as we will see, a good justification of the use of a type A region can be given in the light of some recent results.

III R. A. FISHER'S THEORY OF ESTIMATION[5)]

The problem of estimation of the unknown parameter θ is the problem of finding a function $t(x_1,\ldots,x_n)$ of the observations such that t can be considered in a certain sense as a "good" or "best" estimate of θ. Since the estimate $t(x_1,\ldots,x_n)$ is a random variable, we cannot expect that its value should coincide with that of the unknown parameter, but we will try to choose $t(x_1,\ldots,x_n)$ in such a way as to make as great as possible the probability of the value of t lying as near as possible to the value of the unknown parameter θ.

This is a somewhat vague formulation of the requirement for a "good" or "best" statistical estimate. It can be made precise in different ways. Markoff[6)], for instance, defines the notion of a "best" estimate as follows: A statistic t (we shall call any function of the observations a statistic) is a best estimate of θ if

(1) t is an unbiased estimate of θ, i.e., $E_\theta(t) = \theta$ identically in θ where $E_\theta(t)$ denotes the expected value of t under the assumption that θ is the true value of the parameter.

(2) $E_\theta(t-\theta)^2 \leq E_\theta(t'-\theta)^2$ identically in θ for all t' which satisfy (1).

This definition of a "best estimate" seems to be a reasonable and acceptable one since, in general, the smaller the variance of t the greater is the probability that t will lie in a small

---

5) See references 3 - 6

6) See reference 15, p.344

neighborhood of θ. It should be remarked that although (by virtue of Tshebisheff's inequality) smallness of the variance implies that the probability of t lying in a small neighborhood of θ is small, the converse is not necessarily true. It may happen that a statistic t has a large variance and, nevertheless, the probability of t lying in a small neighborhood of θ is high. This circumstance constitutes some argument against Markoff's definition. A more serious difficulty is, however, the fact that a best estimate in Markoff's sense seldom exists.

R. A. Fisher's theory of estimation is based on the principle of the maximum likelihood. It is assumed that a probability density

$$p(x_1,\ldots,x_n, \theta)$$

exists in the sample space, i.e., for any measurable subset W of the sample space

$$P(W|\theta) = \int_W p(x_1,\ldots,x_n, \theta)\, dx.$$

In particular, the cumulative distribution function is given by

$$F(x_1,\ldots,x_n, \theta) = \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} P(v_1,\ldots,v_n, \theta)\,dv_1,\ldots dv_n.$$

The maximum likelihood estimate $\theta_n(x_1,\ldots,x_n)$ is defined as that value of θ for which $p(x_1,\ldots,x_n,\theta)$ becomes a maximum. Now assume that $X_1,\ldots,X_n$ are n independently distributed random variables each having the same distribution. This can also be expressed by saying that $x_1,\ldots,x_n$ are n independent observations on the same random variable X. The main result of Fisher's theory of estimation can be stated as follows: If $x_1,\ldots,x_n$ are n independent observations (n = 1,..., ad inf.) on the same random variable X and if the distribution of X

satisfies certain conditions (which are not too restrictive and in practical application are frequently fulfilled), then $\hat{\theta}_n$ is an efficient estimate. The definition of an efficient estimate is given as follows:

A sequence $\{t_n\}$ (n = 1,..., ad inf.) of statistics is called an efficient estimate of θ (the subscript n indicates the number of observations of which t_n is a function) if

(1) the limit distribution of $\sqrt{n}\ (t_n - \theta)$ is a normal distribution with zero mean and finite variance, and

(2) for any sequence $\{t'_n\}$ of statistics which satisfies (1)

$$\sigma^2/\sigma'^2 \leqq 1$$

where $\sigma^2 = \lim E_\theta \left[\sqrt{n}\ (t_n - \theta)\right]^2$

and $\sigma'^2 = \lim E_\theta \left[\sqrt{n}\ (t'_n - \theta)\right]^2$

The ratio σ^2/σ'^2 is called the efficiency of $\{t'_n\}$ which is always $\leqq 1$.

Vaguely speaking, in large samples the maximum likelihood estimate has the smallest variance compared with any other statistic which is in the limit normally distributed. The restriction of the comparison to statistics which are in the limit normally distributed seems to be a serious one. However, as recent results show, the maximum likelihood estimate has a much stronger property than efficiency, and it can be considered as a "best" large sample estimate of θ compared even with statistics which are not normally distributed in the limit.[7]

---

7) See reference 20

The question of consistency and limit distribution of the maximum likelihood estimate has been treated by H. Hotelling,7. A complete proof has been given by J. L. Doob, 1.

As an example, let $x_1,\ldots,x_n$ be n independent observations on a normally distributed variate X with unknown mean and unit variance. It can easily be verified that the maximum likelihood estimate of θ is given by

$$\hat{\theta}_n(x_1,\ldots,x_n) = \frac{x_1 +\ldots+ x_n}{n}$$

Let $t_n(x_1,\ldots,x_n)$ be the median of the observations $x_1,\ldots,x_n$. It can be shown that the limit distribution of $\sqrt{n}\,(t_n - \theta)$ is normal with zero mean and variance $\frac{\pi}{2}$. Hence, the efficiency of the median for estimating θ is equal to $\frac{2}{\pi} = 0.6366\ldots$

IV THE THEORY OF CONFIDENCE INTERVALS

The procedure of estimation, as I formulated it here, is also called estimation by a point. For practical applications the estimation by intervals seems to be much more important. That is to say, we have to construct two functions of the observations $\underline{\theta}(E)$ and $\overline{\theta}(E)$, where E denotes a point of the sample space, and we estimate the parameter to be within the interval $\delta(E) = \left[\underline{\theta}(E), \overline{\theta}(E)\right]$. In connection with the theory of interval estimation, R. A. Fisher introduced the notion of fiducial probability and fiducial limits, while Neyman[8] developed the theory of interval estimation based on the classical theory of probability. I shall give here a brief outline of Neyman's theory.

Before the sample has been drawn the point E is a random variable and, therefore, the values of $\underline{\theta}(E)$ and $\overline{\theta}(E)$ are also random variables. Hence, before the sample has been drawn we can speak of the probability that

$$(3) \qquad \underline{\theta}(E) \leq \theta \leq \overline{\theta}(E)$$

even if θ is considered merely as an unknown constant. After the sample has been drawn and we have obtained a particular sample point, say E_0, it does not make sense to speak of the probability that

$$(4) \qquad \underline{\theta}(E_0) \leq \theta \leq \overline{\theta}(E_0),$$

if θ is merely an unknown constant. Each term in the inequality (4) is a fixed constant, and the inequality (4) is either

---

8) See reference 15

right or wrong for those particular constants. It would be proper to talk about the probability of (4) if θ itself could be considered as a random variable having a certain probability distribution, called an a priori probability distribution. In this case we understand by the probability that (4) holds the conditional probability, called also a posteriori probability, under the assumption that $E = E_0$ occurred. If an a priori distribution of θ exists and if it is known then, using Bayes' formula, we can easily calculate the a posteriori probability distribution of θ. However, in practical applications we seldom meet cases where the assumption of the existence of an a priori probability distribution seems to be justified; and even in those rare cases in which the latter assumption can be made, we usually do not know the shape of the a priori probability distribution and this makes the application of Bayes' theorem impossible. For these reasons the theory of interval estimation has to be developed in such a way that its validity should not depend on the existence of an a priori probability distribution. Hence, in this theory we shall speak only of the probability of (3) but never of the probability of (4).

For any relationship R we will denote by $P\left[R \mid \theta\right]$ the probability of R calculated under the assumption that θ is the true value of the parameter.

A pair of functions $\underline{\theta}(E)$ and $\overline{\theta}(E)$ is called *a confidence interval of* θ if

1) $\underline{\theta}(E) \leq \overline{\theta}(E)$ for all points of E

2) $P\left[\underline{\theta}(E) \leq \theta \leq \overline{\theta}(E) \mid \theta\right] = \alpha$ for all values of θ,

where α is a fixed constant called the *confidence coefficient*.

The practical meaning and importance of the notion of the confidence interval is this: If a large number of samples are drawn and if in each case we make the statement that θ is included in the interval $\left[\underline{\theta}(E), \overline{\theta}(E)\right]$, then the relative frequency of correct statements will approximately be equal to α.

In general, there exist infinitely many confidence intervals corresponding to a fixed confidence coefficient α, and we have to set up some principle for choosing from among them. It is obvious that we want the confidence interval corresponding to a fixed confidence coefficient to be as "short" as possible. We have to give a precise definition of the notion "shortest" confidence interval.

A confidence interval $\delta(E) = \left[\underline{\theta}(E), \overline{\theta}(E)\right]$ is called a shortest confidence interval corresponding to the confidence coefficient α if

(a) $P\left[\underline{\theta}(E) \leq \theta \leq \overline{\theta}(E) \mid \theta\right] = \alpha$ and

(b) for any confidence interval $\delta'(E)$ which satisfies (a)

$$P\left[\underline{\theta}(E) \leq \theta' \leq \overline{\theta}(E) \mid \theta''\right] \leq P\left[\underline{\theta}'(E) \leq \theta' \leq \overline{\theta}'(E) \mid \theta''\right]$$

for all values θ' and θ'' of θ.

If a shortest confidence interval exists, it seems to be the most advantageous. Unfortunately, shortest confidence intervals exist only in quite exceptional cases. Therefore, we have to introduce some further principles on which the choice should be based. Such a principle is the principle of unbiasedness.

A confidence interval $\delta(E)$ is called an unbiased confidence interval corresponding to the confidence coefficient α if

$$P\left[\underline{\theta}(E) \leq \theta \leq \overline{\theta}(E) \mid \theta\right] = \alpha$$

and $P\left[\underline{\theta}(E) \leq \theta' \leq \overline{\theta}(E) \mid \theta''\right] \leq \alpha$ for all values θ' and θ''.

A confidence interval $\delta(E)$ is called a shortest unbiased confidence interval corresponding to the confidence coefficient α if $\delta(E)$ is an unbiased confidence interval with the confidence coefficient α and if for any unbiased confidence interval $\delta'(E)$ with the same confidence coefficient, we have

$$P\left[\underline{\theta}(E) \leq \theta' \leq \overline{\theta}(E) \mid \theta''\right] \leq P\left[\underline{\theta}'(E) \leq \theta' \leq \overline{\theta}'(E) \mid \theta''\right]$$

for all values θ' and θ''.

If we accept the principle of unbiasedness, the shortest unbiased confidence interval seems to be the most favorable one. Even shortest unbiased confidence intervals exist only in a restricted, but important, class of cases. If a shortest unbiased confidence interval does not exist, Neyman proposes the use of a third type of confidence interval, which he calls "short unbiased" confidence interval. An unbiased confidence interval $\delta(E)$ with the confidence coefficient α is called a short unbiased confidence interval if

$$\frac{\partial^2}{\partial\theta''^2} P\left[\underline{\theta}(E) \leq \theta' \leq \overline{\theta}(E) \mid \theta''\right]\Big|_{\theta''=\theta'} \leq \frac{\partial^2}{\partial\theta''^2} P\left[\underline{\theta}'(E) \leq \theta' \leq \overline{\theta}'(E) \mid \theta''\right]\Big|_{\theta''=\theta'}$$

for all θ' and for all unbiased confidence intervals $\delta'(E)$ with the confidence coefficient α.

I have discussed only the case of a single unknown parameter. In the case of several unknown parameters some new problems arise, which do not occur in the case of a single parameter. However, I shall not discuss them, since the case of a single parameter already provides a good illustration of the basic ideas of the theories of Fisher, Neyman and Pearson.

V ASYMPTOTICALLY MOST POWERFUL TESTS AND ASYMPTOTICALLY SHORTEST CONFIDENCE INTERVALS[9]

As we have seen, if a uniformly most powerful (unbiased) test and a shortest (unbiased) confidence interval exist, they provide a satisfactory solution of the problem of testing a hypothesis and the problem of interval estimation. Unfortunately, they exist only in a restricted class of cases. As substitutes for them the use of a critical region of type A and a short confidence interval, respectively, have been proposed. The appropriateness of the region of type A seems somewhat doubtful, since we are more interested in the behavior of the power function at values of θ far from the value θ_0 to be tested than at values of θ near to θ_0. Similar objections can be raised to the use of a short confidence interval. Recent investigations show, however, that the situation is much more favorable than appears at first glance. It is shown that the difficulties arising because of the non-existence of uniformly most powerful unbiased tests and shortest unbiased confidence intervals gradually disappear with increasing size of the sample, since so-called asymptotically most powerful unbiased tests and asymptotically shortest unbiased confidence intervals practically always exist.

We shall assume that the observations $x_1, \ldots, x_n$ are n independent observations on the same random variable X whose distribution function involves a single unknown parameter θ. We shall also assume that X has a probability density function,

---

9) See references 17-20

say $f(x,\theta)$. Since in our discussions the number of observations n will not be kept constant, we shall indicate the dimension of the sample space by proper subscripts. For instance, a critical region in the n-dimensional sample space will be denoted by a capital letter with the subscript n. A point of the n-dimensional sample space will be denoted by E_n, and a confidence interval based on n observations by $\delta_n(E_n)$.

For any region U_n denote by $G(U_n)$ the greatest lower bound of $P(U_n|\theta)$. For any pair of regions U_n and T_n denote by $L(U_n,T_n)$ the least upper bound of

$$P\left[U_n(\theta) - P(T_n|\theta)\right].$$

A sequence $\{W_n\}$ (n=1,...,ad inf.) of regions is said to be an asymptotically most powerful test of the hypothesis $\theta = \theta_o$ on the level of significance α if $P(W|\theta_o) = \alpha$ and if for any sequence $\{Z_n\}$ of regions for which $P(Z_n|\theta_o) = \alpha$,

$$\lim_{n\to\infty} \sup L(Z_n,W_n) = 0 \quad \text{holds.}$$

A sequence $\{W_n\}$ (n=1,...,ad inf.) of regions is said to be an asymptotically most powerful unbiased test of the hypothesis $\theta = \theta_o$ on the level of significance α if $P(W_n|\theta_o)=\lim G(W_n)=\alpha$ and if for any sequence $\{Z_n\}$ of regions for which $P(Z_n|\theta_o) = \lim G(Z_n) = \alpha$ the inequality $\limsup L(Z_n,W_n) \leq 0$ holds.

Let $P_n(\theta,\alpha)$ be defined by

$$P_n(\theta,\alpha) = \text{l.u.b. } P(Z_n|\theta)$$

with respect to all regions Z_n for which $P(Z_n|\theta_o) = \alpha$. We will call $P_n(\theta,\alpha)$ the envelope function corresponding to the level of significance α. Similarly let $P_n^*(\theta,\alpha)$ be the least upper bound of $P(Z_n|\theta)$ with respect to all unbiased critical regions Z_n which have the size α. We will call $P_n^*(\theta,\alpha)$ the unbiased envelope function corresponding to the level of significance α.

The two previously given definitions are equivalent to the following two:

A sequence $\{W_n\}$ of regions is said to be an asymptotically most powerful test of the hypothesis $\theta = \theta_0$ on the level of significance α if $P(W_n|\theta_0) = \alpha$ and

$$\lim_{n=\infty} \left\{P_n(\theta,\alpha) - P(W_n|\theta)\right\} = 0$$

uniformly in θ.

A sequence $\{W_n\}$ of regions is said to be an asymptotically most powerful unbiased test of the hypothesis $\theta = \theta_0$ on the level of significance α if $P(W_n|\theta_0) = \alpha$ and

$$\lim_{n=\infty} \left\{P_n^*(\theta,\alpha) - P(W_n|\theta)\right\} = 0$$

uniformly in θ.

Let $\hat{\theta}_n(x_1,\ldots,x_n)$ be the maximum likelihood estimate of θ in the n-dimensional sample space. That is to say, $\hat{\theta}_n$ denotes the value of θ for which the product $\prod_{\alpha=1}^{n} f(x_\alpha,\theta)$ becomes a maximum. Let W_n' be the region defined by the inequality $\sqrt{n}(\hat{\theta}_n - \theta_0) \geq c_n'$, W_n'' defined by the inequality $\sqrt{n}(\hat{\theta}_n - \theta_0) \leq c_n''$ and let W_n be defined by the inequality $|\sqrt{n}(\hat{\theta}_n - \theta_0)| \geq d_n$. The constants d_n, c_n', c_n'' are chosen in such a way that

$$P(W_n'|\theta_0) = P(W_n''|\theta_0) = P(W_n|\theta_0) = \alpha.$$

It has been shown that under certain restrictions on the probability density $f(x,\theta)$ the sequence $\{W_n'\}$ is an asymptotically most powerful test of the hypothesis $\theta = \theta_0$ if θ takes only values $\geq \theta_0$. Similarly $\{W_n''\}$ is an asymptotically most powerful test if θ takes only values $\leq \theta_0$. Finally $\{W_n\}$ is an asymptotically most powerful unbiased test if θ can take any real value.

There are also other asymptotically most powerful tests. Let W_n' be the region defined by the inequality

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} \frac{\partial}{\partial\theta} \log f(x_\alpha, \theta_0) \geq c_n' ,$$

W_n'' defined by the inequality

$$\frac{1}{\sqrt{n}} \sum_{\alpha} \frac{\partial}{\partial\theta} \log f(x_\alpha, \theta_0) \leq c_n'' ,$$

and W_n defined by the inequality

$$\left| \frac{1}{\sqrt{n}} \frac{\partial}{\partial\theta} \sum_{\alpha} \log f(x_\alpha, \theta_0) \right| \geq c_n$$

where the constants c_n, c_n' and c_n'' are chosen in such a way that

$$P(W_n' | \theta_0) = P(W_n'' | \theta_0) = P(W_n | \theta_0) = \alpha.$$

Then $\{W_n'\}$ is an asymptotically most powerful test of the hypothesis $\theta = \theta_0$ if θ takes only values $\geq \theta_0$. Similarly, $\{W_n''\}$ is an asymptotically most powerful test if θ takes only values $\leq \theta_0$. Finally $\{W_n\}$ is an asymptotically most powerful unbiased test if θ can take any real value.

The sequence $\{A_n(\theta_0)\}$ is an asymptotically most powerful unbiased test of the hypothesis $\theta = \theta_0$, where $A_n(\theta_0)$ denotes the critical region of type A for testing the hypothesis $\theta = \theta_0$.

Since there are many asymptotically most powerful tests, the question arises whether they are all equally good or whether one can be preferred to another. It is clear that if $\{W_n\}$ and $\{W_n'\}$ are two asymptotically most powerful unbiased tests, then for sufficiently large n they are equally good. In fact, for sufficiently large n both power functions $P(W_n | \theta)$ and

$P(W_n'|\theta)$ are in a small neighborhood of $P_n(\theta,\alpha)$ $\left[P_n^*(\theta,\alpha)\right]$. However, they may behave differently in the sense that with increasing n one power function, say $P(W_n|\theta)$ approaches the envelope function faster than $P(W_n'|\theta)$ does. In such a case it seems preferable to use W_n, especially if the sample is only moderately large. If the sample is so large that both power functions are in a small neighborhood of the envelope function, then it is immaterial whether we use W_n or W_n'.

These considerations lead to the idea that it is preferable to use that asymptotically most powerful (unbiased) test $\{W_n\}$ for which the approach of $P(W_n|\theta)$ to the envelope function is, in a certain sense, fastest.

A region W_n is called a most stringent test of size α for testing the hypothesis $\theta = \theta_0$ if $P(W_n|\theta_0) = \alpha$ and

$$\underset{\theta}{\text{l.u.b.}}\left[P_n(\theta,\alpha)-P(W_n|\theta)\right] \qquad \underset{\theta}{\text{l.u.b.}}\left[P_n(\theta,\alpha)-P(Z_n|\theta)\right]$$

for all Z_n for which $P(Z_n|\theta_0) = \alpha$. The abbreviation $\underset{\theta}{\text{l.u.b.}}$ means "least upper bound with respect to θ."

If W_n is for each n a most stringent test, its power function will approach the envelope function, in a certain sense, faster than any other power function. It seems, therefore, desirable to use a most stringent test. A region of type A is not exactly a most stringent test, but probably it is quite near to it (this question has yet to be investigated), and this would provide a very good justification for the use of a type A region. The mathematical difficulties in finding explicitly a most stringent test are considerable.

Let $\delta_n(E_n) = \left[\underline{\theta}_n(E_n), \overline{\theta}_n(E_n)\right]$ be an interval function and denote by $P\left[\delta_n(E_n)\ C\theta' \mid \theta''\right]$ the probability that $\delta_n(E_n)$ will cover θ' under the assumption that θ'' is the true value of the parameter.

A sequence of interval functions $\left\{\delta_n(E_n)\right\}$ (n=1,2,..., ad inf) is called an asymptotically shortest confidence interval of θ if the following two conditions are fulfilled:

(a) $P\left[\delta_n(E_n)\ C\theta \mid \theta\right] = \alpha$ for all values of θ

(b) For any sequence of interval functions $\left\{\delta'_n(E_n)\right\}$ (n=1,2,..., ad inf.) which satisfies (a), the least upper bound of

$$P\left[\delta_n(E_n)\ C\theta' \mid \theta''\right] - P\left[\delta'_n(E_n)\ C\theta' \mid \theta''\right]$$

with respect to θ' and θ'' converges to zero with $n \to \infty$.

A sequence of interval functions $\left\{\delta_n(E_n)\right\}$ (n=1,2,..., ad inf) is called an asymptotically shortest unbiased confidence interval of θ if the following three conditions are fulfilled:

(a) $P\left[\delta_n(E_n)\ C\theta \mid \theta\right] = \alpha$ for all values of θ

(b) The least upper bound of $P\left[\delta_n(E_n)\ C\theta' \mid \theta''\right]$ with respect to θ' and θ'' converges to α with $n \to \infty$

(c) For any sequence of interval functions $\left\{\delta'_n(E_n)\right\}$ which satisfies the conditions (a) and (b), the least upper bound of

$$P\left[\delta_n(E_n)\ C\theta' \mid \theta''\right] - P\left[\delta'_n(E_n)\ C\theta' \mid \theta''\right]$$

with respect to θ' and θ'', converges to zero with $n \to \infty$.

Let $C_n(\theta)$ be a positive function of θ such that the probability that $\left|\frac{1}{\sqrt{n}} \sum_{\beta} \frac{\partial}{\partial\theta} \log f(x_\beta, \theta)\right| \leq C_n(\theta)$ is equal to α

constant α under the assumption that θ is the true value of the parameter. Denote by $\underline{\theta}(E_n)$ the root in θ of the equation

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \sum_{\beta} \log f(x_\beta, \theta) = C_n(\theta)$$ and by $\overline{\theta}(E_n)$ the root of

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \sum_{\beta} \log f(x_\beta, \theta) = -C_n(\theta).$$ It has been shown that under some restrictions on $f(x,\theta)$ the interval $\delta(E_n) = \left[\underline{\theta}(E_n), \overline{\theta}(E_n)\right]$ is an asymptotically shortest unbiased confidence interval of θ corresponding to the confidence coefficient α. This confidence interval is identical with that given by Wilks[10]. The definition of a shortest confidence interval underlying Wilks' investigations is somewhat different from that of Neyman's, which has been used here. According to Wilks, a confidence interval $\delta(E)$ is called shortest in the average if the expectation of the length of $\delta(E)$ is a minimum. The main result obtained by Wilks can be formulated as follows: The confidence interval in question is asymptotically shortest in the average compared with all confidence intervals the endpoints of which are roots of an equation of the following type:

$$\sum_{\beta} h(x_\beta, \theta) = \pm C_n(\theta).$$

In the present investigation such a restriction is not made. The confidence interval in consideration is shown to be asymptotically shortest compared with any unbiased confidence interval.

Now let $C_n(\theta)$ be a positive function of θ such that the probability that $\left|\hat{\theta}_n - \theta\right| \leq C_n(\theta)$ is equal to a constant α under

---

10) See reference 22

the assumption that θ is the true value of the parameter. Denote by $\underline{\theta}(E_n)$ the root in θ of the equation $\hat{\theta}_n - \theta = C_n(\theta)$ and by $\overline{\theta}(E_n)$ the root of $\hat{\theta}_n - \theta = -C_n(\theta)$. Consider the interval $\delta(E_n) = \left[\underline{\theta}(E_n), \overline{\theta}(E_n)\right]$. Under some restrictions on the density $f(x,\theta)$, it can be shown that $\delta(E_n)$ is an asymptotically shortest unbiased confidence interval.

This is a much stronger property of the maximum likelihood estimate than its efficiency and gives a justification of the use of the maximum likelihood estimate also in the light of Neyman's theory of estimation.

VI OUTLINE OF A GENERAL THEORY OF STATISTICAL INFERENCE

The theories of Fisher, Neyman and Pearson are restricted in two respects. First, they consider only the problem of testing a hypothesis and that of estimation by point or interval. The second restriction is that only the case in which Ω is a k-parameter family of distribution functions is investigated. Both restrictions are serious from the point of view of applications.

There are many important statistical problems which are neither problems of testing a hypothesis, nor problems of estimation. We have already given such an example in Section 1. As a further illustration, let us consider the following case: Let $X_1,\ldots,X_p$ be p independently and normally distributed random variables with unit variances and unknown means $\theta_1,\ldots\theta_p$. Furthermore, let $x_{i1},\ldots,x_{in}$ be n independent observations on X_i $(i = 1,2,\ldots,p)$. Suppose we test the hypothesis that $\theta_1 = \ldots = \theta_p = 0$, and decide to reject this hypothesis on the basis of the pn observations $x_{i\alpha}$ $(\alpha = 1,2,\ldots,n;\ i = 1,2,\ldots,p)$. In such cases we are usually interested in knowing which mean values are not zero, i.e., we wish to subdivide the set of p mean values $\theta_1,\ldots,\theta_p$ into two subsets, such that one of them contains the mean values which are zero and the other the mean values which are not zero. This subdivision has to be done, of course, on the basis of the pn observations $x_{i\alpha}$. More precisely, we have to deal with the following statistical problem: There exist 2^p different subsets of the set $(\theta_1,\ldots,\theta_p)$. Denote these subsets by $\omega_1,\ldots,\omega_{2^p}$, respectively. Let H_k

($k = 1,\ldots,2^p$) be the hypothesis that the mean values contained in the set ω_k are equal to zero and all other mean values are unequal to zero. On the basis of the pn observations we have to decide which hypothesis H_k from the set of the 2^p possible hypotheses should be accepted. This problem cannot be considered as a problem of testing a hypothesis nor a problem of estimation.

A similar problem arises if we wish to classify a set of regression coefficients into the class of non-zero and the class of zero regression coefficients. In problems of regression we often take it for granted that the regression in question is a polynomial and we have to determine on the basis of the observations the degree of the polynomial to be fitted. That is to say, we have to decide on the basis of the observations which hypothesis of the sequence of hypotheses H_1, H_2, $H_3,\ldots,$ $H_n,\ldots$ should be accepted. The symbol H_n ($n = 1,2,\ldots$) denotes the hypothesis that the regression is a polynomial of n-th degree. These examples illustrate sufficiently the necessity of the extension of the theory of statistical inference to the general case as formulated in Section 1.

The case in which Ω cannot be represented as a k-parameter family of distribution functions is quite important. As an illustration, consider the following problem: Let $(x_1,y_1),\ldots$ (x_n,y_n) be n independent pairs of observations on a pair (X,Y) of random variables. Suppose we wish to test the hypothesis that X and Y are independently distributed and we do not have any a priori knowledge about the joint distribution of X and Y. In this case Ω consists of all distribution functions

$F(x_1,y_1,\ldots,x_n,y_n)$ which can be written in the form

$$F(x_1,y_1,\ldots) = \overline{\Phi}(x_1,y_1) \ldots \overline{\Phi}(x_n,y_n)$$

where $\overline{\Phi}$ may be an arbitrary function. The subclass ω consists of all distribution functions $F(x_1,y_1,\ldots,x_n,y_n)$ which can be written in the form

$$F(x_1,y_1,\ldots,x_n,y_n) = \varphi(x_1)\psi(y_1)\varphi(x_2)\psi(y_2)\ldots\varphi(x_n)\psi(y_n).$$

Hence, Ω cannot be represented as a k-parameter family of functions.

The problem given above as an illustration has been treated by H. Hotelling and Margaret Pabst (see reference 8). Another problem, where Ω is the class of all continuous distributions, has been considered in paper (see reference 21). We shall give here an outline of a theory of statistical inference dealing with the following general problem[11]:

Let $X_1,\ldots,X_n$ be a set of n random variables. It is known that the joint probability distribution function $F(x_1,\ldots,x_n)$ of $X_1,\ldots,X_n$ is an element of a certain class Ω of distribution functions. Let S be a system of subclasses of Ω. For each element ω of S denote by H_ω the hypothesis that the true distribution $F(x_1,\ldots,x_n)$ of $X_1,\ldots,X_n$ is an element of ω. Denote by H_S the system of all hypotheses corresponding to all elements of S. Let x_i be the observed value of X_i $(i=1,\ldots,n)$. We have to decide by means of the observed sample point $E_n = (x_1,\ldots,x_n)$ which hypothesis of the system H_S of hypotheses should be accepted. That is to say, for each hypothesis H_ω we have to determine a region of acceptance M_ω in the n-dimensional sample space. The hypothesis H_ω will be accepted

---

11) This theory has been developed in reference 16 for the case that Ω is a k-parameter family

if and only if the sample point falls in the region M_ω. The regions M_ω and $M_{\omega'}$ are, of course, disjoint for $\omega \neq \omega'$. Furthermore, $\sum_\omega M_\omega$ is equal to the whole sample space. The statistical problem is that of the proper choice of the system M_S of the regions of acceptance.

The choice of the system M_S of regions of acceptance is equivalent to the choice of a function $\omega(E_n)$ defined over all points E_n of the sample space. The value of the function $\omega(E_n)$ is an element of S determined as follows: Since the elements of M_S are disjoint and since $\sum_\omega M_\omega$ is equal to the whole sample space, for each point E_n there exists exactly one element ω of S such that E_n is contained in M_ω. The value of the function $\omega(E_n)$ is that element ω of S for which E_n is an element of M_ω. Hence, we can replace M_S by the function $\omega(E_n)$ and for each sample point E_n we decide to accept the hypothesis $H_\omega(E_n)$. We will call $\omega(E_n)$ the statistical decision function. Hence, the statistical problem is that of choosing the statistical decision function $\omega(E_n)$.

The choice of $\omega(E_n)$ will essentially be affected by the relative importance of the different possible errors we may commit. We commit an error whenever we accept a hypothesis H_ω and the true distribution is not an element of ω. We introduce a weight function for the possible errors. The weight function $w[F,\omega]$ is a real valued non-negative function defined for all elements F of Ω and all elements ω of S, expressing the relative importance of the error committed by accepting H_ω when F is true. If F is an element of ω then $w[F,\omega] = 0$, otherwise $w[F,\omega] > 0$. The question as to how the form of the weight function $w[F,\omega]$ should be chosen is not a mathematical nor statistical

one. The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors and this will depend on the special purposes of his investigation. If this is done, we shall in general be able to give a more satisfactory answer to the question as to how the statistical decision function should be chosen. In many cases, especially in statistical questions concerning industrial production, we are able to express the importance of an error in monetary terms, that is, we can express the loss caused by the error considered in terms of money. We shall also say that $w[F,\omega]$ is the loss caused by accepting H_ω when F is true.

Suppose that we make our decisions according to a statistical decision function $\omega(E_n)$, and that the true distribution is the element $F(x_1,\ldots,x_n)$ of Ω. Then the expected value of the loss is obviously given by the Stieltjes integral

$$(5) \qquad \int_{M_n} w[F,\omega(E_n)]\,dF(x_1,\ldots,x_n) = r[F],$$

where the integration is to be taken over the whole sample space M_n. We shall call the expression (5) the risk of accepting a false hypothesis when F is the true distribution function. Since we do not know the true distribution F we shall have to study the risk $r[F]$ as a function of F. We shall call this function the <u>risk function</u>. Hence, the risk function is defined over all elements F of Ω. The form of the risk function depends on the statistical decision function $\omega(E_n)$ and on the weight function $w[F,\omega]$. In order to express this fact, we shall denote the risk function associated with the statistical decision function $\omega(E_n)$ and the weight function $w[F,\omega]$ also by

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\}$$

We introduce the following definitions:

Definition 1. Denote by $\omega(E_n)$ and $\omega'(E_n)$ two statistical decision functions for the same system H_g of hypotheses. We shall say that $\omega(E_n)$ and $\omega'(E_n)$ are equivalent relative to the weight $w[F,\omega]$ if the risk function $r\left\{F|\omega(E_n),\ w[F,\omega]\right\}$ is identically equal to the risk function $r\left\{F|\omega'(E_n), w[F,\omega]\right\}$ i.e., for any element F of Ω we have

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\} = r\left\{F|\omega'(E_n),\ w[F,\omega]\right\}.$$

Definition 2. Denote by $\omega(E_n)$ and $\omega'(E_n)$ two statistical decision functions for the same system H_g of hypotheses. We shall say that $\omega(E_n)$ is uniformly better than $\omega'(E_n)$ relative to the weight function $w[F,\omega]$ if $\omega(E_n)$ and $\omega'(E_n)$ are not equivalent and for each element F of Ω we have

$$r\left\{F|\omega(E_n),\ w[F,\omega]\right\} \leq r\left\{F|\omega'(E_n),\ w[F,\omega]\right\}.$$

Definition 3. A statistical decision function $\omega(E_n)$ is said to be admissible relative to the weight function $w[F,\omega]$ if no uniformly better statistical decision function exists relative to the weight function considered.

First principle for the choice of the statistical decision function. We choose a statistical decision function which is admissible relative to the weight function considered.

There can scarcely be given any argument against the acceptance of the above principle for the selection of $\omega(E_n)$. However, this principle does not lead in general to a unique solution. There exist in general many admissible statistical decision functions. We need a second principle for the choice of a best admissible decision function.

The choice between two admissible decision functions $\omega(E_n)$ and $\omega'(E_n)$ may be affected by the degree of our a priori confidence in the truth of the different elements of Ω. Suppose, for instance, that for a certain element F_1 of Ω we have

$$r\left\{F_1|\omega(E_n), w[F,\omega]\right\} < r\left\{F_2|\omega'(E_n), w[F,\omega]\right\}$$

for another element F_2 of Ω we have

$$r\left\{F_2|\omega(E_n), w[F,\omega]\right\} > r\left\{F_2|\omega'(E_n), w[F,\omega]\right\}$$

and for any other element $F \neq F_1, \neq F_2$ we have

$$r\left\{F|\omega(E_n), w[F,\omega]\right\} = r\left\{F|\omega'(E_n), w[F,\omega]\right\}.$$

If we have much greater a priori confidence in the truth of F_1 than in that of F_2, we will probably prefer $\omega(E_n)$ to $\omega'(E_n)$. On the other hand, if we think a priori that F_2 is more likely to be true than F_1, we may prefer $\omega'(E_n)$ to $\omega(E_n)$.

Suppose we can express our a priori degree of confidence by a non-negative additive set function $\rho(\eta)$ defined over a certain system of subsets η of Ω, where $\rho(\Omega) = 1$. That is to say the value of $\rho(\eta)$ expresses the degree of our a priori belief that the true distribution is an element of the subset η. In such a case it seems very reasonable to consider a decision function $\omega^*(E_n)$ as "best" if the value of the integral

$$\int_\Omega r\left\{F|\omega(E_n), w[F,\omega]\right\} d\rho$$

becomes a minimum for $\omega(E_n) = \omega^*(E_n)$. That is, we consider a decision function $\omega^*(E_n)$ as "best" if it minimizes a certain weighted average of the risk function.

However, it is doubtful that a set function expressing our a priori degree of belief can meaningfully be constructed. Therefore, we prefer to formulate the notion of a "best" decision function independently of such considerations.

Denote by $r\left\{\omega(E_n),\ w\left[F,\omega\right]\right\}$ the least upper bound of $r\left\{F|\omega(E_n),\ w\left[F,\omega\right]\right\}$ with respect to F, where F may be any element of Ω.

Definition 4. A decision function $\omega^*(E_n)$ is said to be a "best" decision function if $r\left\{\omega(E_n),\ w\left[F,\omega\right]\right\}$ becomes a minimum for $\omega(E_n) = \omega^*(E_n)$. (The weight function $w\left[F,\omega\right]$ is considered fixed.)

This definition of a "best" decision function seems to be a very reasonable one, although it is not the only possible one. One could reasonably define a decision function as "best" if it minimizes a certain weighted average of the risk function. However, there are certain properties of the "best" decision function according to definition 4, which seem to justify the use of that definition. One of the most important properties of a "best" decision function in the sense of definition 4 is that the risk function is a constant, i.e., it has the same value for all elements F of Ω. This has been shown in the case that Ω is a k-parameter family of distributions, and the weight function $w\left[F,\omega\right]$ and the distribution functions F satisfy certain restrictive conditions. The constancy of the risk function seems to be very desirable from the point of view of applications since this property makes it possible to evaluate the exact magnitude of the risk associated with the statistical decision. In the theory of confidence intervals the confidence coefficient, α, i.e., the probability that the confidence interval will cover the unknown parameter, is independent of the value of the unknown parameter. This fact, which is considered to be of basic importance in the theory of interval-estimation,

is analogous to the constancy of the risk function in our general theory since 1-α can be considered in a certain sense as the risk associated with the interval estimation. (The quantity 1-α is exactly equal to the risk in the sense of our definition, if the weight function takes only the values 0 and 1.)

Finally, I should like to make some remarks about the relationship of the general theory as outlined here, to the particular theory of uniformly most powerful and asymptotically most powerful tests which were discussed before. In the case of testing the simple hypothesis that the unknown distribution $F(x_1,\ldots,x_n)$ is equal to a particular distribution $F_o(x_1,\ldots x_n)$, the system S of subsets of Ω consists only of two elements ω_1 and ω_2 where ω_1 contains the single element F_o and ω_2 is the complement of ω_1 in Ω. Hence, the decision function $\omega(E_n)$ can assume merely the values ω_1 and ω_2. Let M_{ω_1} be the subset of the sample space consisting of the points E_n for which $\omega(E_n)=\omega_1$ and let M_{ω_2} be the set of points E_n for which $\omega(E_n)=\omega_2$. The set M_{ω_2} is the complement of M_{ω_1} in the sample space. Obviously the set M_{ω_2} is the critical region, in the sense of the Neyman-Pearson theory. It is easy to see that if for any $\alpha(0<\alpha<1)$ a uniformly best critical region of size α for testing $F = F_o$ exists, then for any arbitrary weight function and for any admissible (see definition 3) decision function $\omega(E_n)$, the set M_{ω_2} will be a uniformly best critical region. In particular, the set M_{ω_2} corresponding to the "best" decision function (see definition 4) will be a uniformly best critical region. Hence, the form of the weight function affects merely the size of the region M_{ω_2} associated with the "best" decision function $\omega(E_n)$,

but it will always be a uniformly best critical region in the sense of the Neyman-Pearson theory. Similar considerations hold concerning asymptotically most powerful tests. Let the sequence $\{W_n\}$ (n=1,2,...,ad inf.) of critical regions be an asymptotically most powerful test for testing the simple hypothesis $F = F_0$. Then for sufficiently large n the region W_n is practically a uniformly best critical region and, therefore, it will be an excellent approximation to the region which is "best" in the sense of definition 4 irrespective of the shape of the weight function of errors.

As we have seen, for building up a general theory of statistical inference, the following three steps have to be made:

1. Formulation of the general problem of statistical inference.
2. Definition of the "best" procedure for making statistical decisions, i.e., definition of the "best" statistical decision function.
3. Solution of the mathematical problem of calculating the "best" statistical decision function.

The problem of statistical inference, as we have formulated it here, seems to be sufficiently broad to cover the problems in practical applications. The second step will always be, to a certain extent, arbitrary. The definition of "best" decision function given here seems to be a satisfactory one. Moreover, under certain restrictive conditions it has the important property that the risk function associated with the "best" decision function is constant, i.e., it has the same value for all elements of Ω. However, there may be other definitions of a

"best" decision function worth investigating. Decision functions which minimize a certain average of the risk function may be of special interest. Concerning step 3, there are many mathematical problems as yet unsolved.

REFERENCES

1 J.L.Doob — PROBABILITY AND STATISTICS, Transactions of the American Mathematical Society, Vol. 36, pp. 759-775

2 J.L.Doob — STATISTICAL ESTIMATION, Transactions of the American Mathematical Society, Vol. 39, pp. 410-421

3 R.A.Fisher — ON THE MATHEMATICAL FOUNDATION OF THEORETICAL STATISTICS, Philosophical Transactions of the Royal Society of London, Series A, Vol. 222, (1921), pp. 309-368

4 R.A.Fisher — THE THEORY OF STATISTICAL ESTIMATION, Proceedings of the Cambridge Philosophical Society, Vol. 22 (1925), pp. 700-725

5 R.A.Fisher — THE FIDUCIAL ARGUMENT IN STATISTICAL INFERENCE, Annals of Eugenics, Vol. 6 (1935), pp.391-398

6 R.A.Fisher — STATISTICAL THEORY OF ESTIMATION, University of Calcutta, 1938

7 H.Hotelling — THE CONSISTENCY AND ULTIMATE DISTRIBUTION OF OPTIMUM STATISTICS, Transactions of the American Mathematical Society, Vol. 32, pp. 847-859

8 H.Hotelling and M. Pabst — RANK CORRELATION AND TESTS OF SIGNIFICANCE INVOLVING NO ASSUMPTION OF NORMALITY, Annals of Mathematical Statistics, Vol. 7, pp. 29-43

9 A.Kolmogoroff — GRUNDBEGRIFFE DER WAHRSCHEINLICHKEITSRECHNUNG, Ergebnisse der Mathematik, Vol. 2, No. 3

10 R.v.Mises — WAHRSCHEINLICHKEITSRECHNUNG UND IHRE ANWENDUNG IN DER STATISTIK UND THEORETISCHEN PHYSIK, Deuticke, Leipzig, 1931

11 R.v.Mises — PROBABILITY, STATISTICS AND TRUTH, William Hodge and Company, London, 1939

12 J.Neyman and E.S.Pearson — ON THE PROBLEM OF THE MOST EFFICIENT TESTS OF STATISTICAL HYPOTHESES, Philosophical Transactions of the Royal Society of London, Series A, Vol. 231 (1933) p. 289

13 J. Neyman and E.S.Pearson — STATISTICAL RESEARCH MEMOIRES, Vol. I, University College, London, 1936, pp. 1-37

14	J. Neyman and E.S.Pearson	STATISTICAL RESEARCH MEMOIRES, Vol. II, University College, London, 1938
15	J. Neyman	OUTLINE OF A THEORY OF STATISTICAL ESTIMATION BASED ON THE CLASSICAL THEORY OF PROBABILITY, Philosophical Transactions of the Royal Society of London, Series A, Vol. 236, pp. 333-380, 1937
16	A.Wald	CONTRIBUTIONS TO THE THEORY OF STATISTICAL ESTIMATION AND TESTING HYPOTHESES, Annals of Mathematical Statistics, Vol. 10, December, 1939
17	A.Wald	ASYMPTOTICALLY MOST POWERFUL TESTS OF STATISTICAL HYPOTHESES, Annals of Mathematical Statistics, Vol. 12, March, 1941
18	A.Wald	SOME EXAMPLES OF ASYMPTOTICALLY MOST POWERFUL TESTS, Annals of Mathematical Statistics Vol. 12, December, 1941
19	A.Wald	A NEW FOUNDATION OF THE METHOD OF MAXIMUM LIKELIHOOD, Report of the Cowles Commission Conference at Colorado Springs, July, 1940
20	A.Wald	ASYMPTOTICALLY SHORTEST CONFIDENCE INTERVALS, paper presented at the meeting of the American Mathematical Society at Hanover, September, 1940. Abstract published in the Bulletin of the American Mathematical Society, Vol. 46, November, 1940
21	A.Wald and J.Wolfowitz	CONFIDENCE LIMITS FOR CONTINUOUS DISTRIBUTION FUNCTIONS, Annals of Mathematical Statistics, Vol. 10, pp. 105-118
22	S.S.Wilks	SHORTEST AVERAGE CONFIDENCE INTERVALS FROM LARGE SAMPLES, Annals of Mathematical Statistics, Vol. 9, pp. 272-280
23	S.S.Wilks and J.F.Daly	AN OPTIMUM PROPERTY OF CONFIDENCE REGIONS ASSOCIATED WITH THE LIKELIHOOD FUNCTION, Annals of Mathematical Statistics, Vol.10, pp. 225-235